Instance scoring via distillation of multiple instance classifiers for interpretable digital pathology

Fumiya Inaba[1], Mira Keyes[2], Calum MacAulay[1], and Martial Guillaud[1]

[1]Basic and Translational Research, BC Cancer Research Institute, 675 W10th Avenue, Vancouver, BC, Canada, V5Z 0B4
[2]Radiation Oncology, BC Cancer Agency, 600 W 10th Ave, Vancouver, BC, Canada, V5Z 4E6

ABSTRACT

Multiple instance learning (MIL) has become a popular approach in computational pathology and whole slide image (WSI) analysis for their weakly-supervised nature. The introduction of attention-based instance pooling in particular has enabled enhanced interpretation of both model decision-making and the underlying data by leveraging attention weights. However, the interpretation of the attention weights, specifically its indication of instance classes, have been contested. We demonstrate that noise or heterogeneity in bagged data can require MIL classifiers to predict bag class based on prevalence of the positive instance class, as opposed to its presence; altering the behavior of the attention mechanism and likely contributing to the aforementioned discrepancies of attention weight interpretation. Here, we introduce an approach to identify and score instances which contribute to a positive bag label, robust against altered attention behavior, in two discreet settings. First, we elucidate how the behavior of attention-based pooling is altered, using the MNIST dataset, where bagged datasets are generated with a different threshold (τ) of positive class instances defining the bag label. While maintaining a high bag-level classification score, the distribution of attention weights between positive and negative changed with τ, where the mechanism attended more to positive instances for lower values of τ, but favored negative instances for high values of τ. We also apply our method to an *in-house* dataset of prostate cancer nuclei to predict the aggressiveness of the disease, and demonstrate how our method may be used to identify a subgroup of nuclei more highly associated with aggressiveness.

Keywords: Multiple instance learning, prostate cancer, interpretability, attention mechanisms

1. INTRODUCTION

Artificial intelligence and machine learning (ML) are becoming popular approaches to analyze whole slide images (WSI) of histological sections, because of their high throughput, consistency and accuracy.[1,2] These approaches are particularly useful in tasks such as prognosis of cancer patients, where traditional methods may fail to capture the heterogeneity between patients. Here, we specifically discuss applications in prostate cancer (PCa), which remains a large health burden as the most commonly diagnosed, and second leading cause of cancer-related deaths for men in Canada.[3] The current clinical standard of risk assessment of PCa patients is the D'Amico risk stratification system, which groups patients into a low, intermediate and high risk group based on serum prostate-specific antigen (PSA) levels, tissue dedifferentiation measured by Gleason score and tumor stage.[4,5] However, PCa is a highly heterogeneous disease and the outcomes of PCa patients are diverse, even within the same defined risk group.[6,7] This has garnered interest in more granular prognostic methods which can measure this heterogeneity, and inform clinical decisions for more personalized options. While other prognostic methods and nomograms with additional clinical variables have improved the prognostic ability, clinical variables may still be limited in the cell-level or molecular detail it can measure. Characteristics such as Gleason scores capture more detail, such as tissue architecture, but are subject to interobserver variability. Thus, quantitative analysis of WSIs is an attractive choice as they are collected routinely, and with specialized stains such as the DNA-stoichiometric Feulgen-thionin stain, large molecular aberrations of the nucleus can be observed on a holistic

Further author information: (Send correspondence to Fumiya Inaba)
Fumiya Inaba: E-mail: finaba@bccrc.ca

Medical Imaging 2026: Digital and Computational Pathology, edited by John E. Tomaszewski
Scott Doyle, Proc. of SPIE Vol. 13932, 1393205 · 2026 Published by SPIE
1605-7422 · doi: 10.1117/12.3087751

level.[8,9] However, detailed supervised annotations for each instance, such as an image patch or cell nucleus, requires a high degree of clinical expertise, and is labor intensive. Weakly-supervised learning methods, such as multiple instance learning (MIL) relax the need for supervised labels, only requiring labels for groups of instances. Additionally, the use of attention mechanisms for MIL has become popular as it enables a deeper interpretation of how individual instances influence the prediction and deeper insights of underlying tissue.[10] For instance, Cai et al. generate a novel attribute score by decoupling spatial attention from instance classification, and Ko et al. leverages attention over clusters of image patches to capture semantic tissue type connotations.[11,12] More notably, MIL-based systems are now being approved by the US Food and Drug Administration (FDA). For instance, Paige Prostate Suite is based on MIL to assist pathologists in WSI review,[13,14] and more recently, Artera AI Prostate, which predicts long-term outcomes for patients with localized PCa,[15,16] was approved for use. While the exact extent of use and reliance remains unclear, clinically approved ML methods demonstrate clinical utility and promise in the future of healthcare.

As these methods get integrated closely with clinical workflows, interpretability of model behavior and decision-making become crucial. The pursuit of interpretable MIL decision-making has popularized the notion of making models interpretable by design, where specific modules have dedicated functionality. For example, Prototypical MIL is designed with a prototype discovery module and slide embedding module, Cluster-aware attention-based MIL has modules for clustering patch-embeddings, Additive MIL aims to achieve intrinsic interpretability among many others.[12,17,18] Such design considerations are crucial for interpretability at scale, yet integration of multiple modules may increase overall complexity, trading off interpretability that comes with simpler systems. While most works alter model design to enforce intended behavior, we argue that the nature of the problem itself can change model behavior, and therefore its interpretation. One illustrative case is the interpretation of the attention weights generated by ABMIL.[10] Ilse et al. demonstrates high attention weights corresponding to positive class instances, while more recent works argue that attention weights alone are not a reliable indicator of instance class.[11] Further, we argue that the behavior of the attention mechanism can change along with the nature of the dataset. The MIL problem is built on the assumption that negatively labeled bags (in a binary classification setting) are completely devoid of instances from the positive class. However, noisy features or inherent biological heterogeneity in pathology may lead to the presence of positive class instances in negative bags in smaller prevalence's than in positive bags.

Here, we take a closer look at how the ABMIL model identifies predictive instances and introduce a method to assign a corresponding class probability to each instance with no additional components. We study the behavior of attention mechanisms of ABMIL in two discrete settings. In the first setting, we generate a collection of datasets in which images of hand-written digits from the MNIST dataset are bagged and labeled according to the prevalence of a positive class instance. In the second, we apply our framework to an *in-house* dataset of PCa patients whose pre-consultation biopsies were stained with Feulgen-thionin. Cell nuclei phenotypes are quantified using a combination of morphological, stain intensity, and image texture features referred to as *Large-scale DNA Organization (LDO)*.[19,20] We leverage LDO as opposed to patches of H&E images for improved biological interpretability, as LDO measures a holistic representation of a cell's molecular, genetic and epigenetic state through the analysis of chromatin distribution.[9,21] We and others have also previously demonstrated the utility of LDO as a predictive and prognostic marker in various cancer contexts.[22–25] In summary, we contribute (1) the description of prevalence-based MIL (prev-MIL) and how attention mechanism behavior can change with instance class prevalence and (2) a method to infer instance class and probability from ABMIL without any additional components.

2. METHODS

2.1 Prevalence-based Multiple Instance Learning

Multiple instance learning (MIL) describes a task wherein a bag of k instances, $X = \{x_1, ..., x_k\}$ with a binary bag label Y_{bag} and binary instance-wise labels $\{y_1, ..., y_k\}$. Traditionally, the binary MIL problem states that a bag is assigned a positive label if there is at least one instance of the positive class in the bag ($Y_{\text{bag}} = 0,$ iff $\sum_{i=0}^{k} y_i = 0$). However, in real-world datasets, particularly in pathology, feature noise and heterogeneity may lead to positive

instances being found in both positive and negatively labeled bags. We thus extend the MIL formulation for use in pathology as:

$$Y_{bag} = \begin{cases} 0 & \text{iff } \frac{\sum_k y_k}{k} < \tau \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

where τ is the proportion of positive instances that a bag must contain to be considered positive. While τ is unknown in most settings, the traditional MIL problem can be described as the case where $\tau = 0$.

2.1.1 Attention-based Multiple Instance Learning

We utilize the implementation of ABMIL introduced in Ref. 11 wherein bag-level predictions are generated by an attention-weighted average of instance logits of the ABMIL classification head as shown in Figure 1C. This approach is equivalent to the original implementation in Ref. 10 provided that no non-linear activation functions are applied to the instance logits. The linear classification head of the ABMIL can be described as:

$$\hat{Y}_{\text{bag}} = \sigma(b + h_{\text{bag}} W^\top) = \sigma(b + (\sum_{i=1}^{k} a_i h_i) W^\top) = \sigma(\sum_{i=1}^{k} a_i z_i) \tag{2}$$

where $\hat{Y}_{\text{bag}}$ is the bag-level class probability, σ is the sigmoid non-linear activation function, $b \in \mathbb{R}^{1\times 1}, W \in \mathbb{R}^{1\times D}$ are learned parameters for instance or bag-level embeddings $h_{i|bag} \in \mathbb{R}^{1\times D}$, a_i is the softmax-normalized instance attention weight and z_i is the instance logit.

2.2 Datasets and Experiments

Two discrete experiment settings are set to elucidate the behavior of the attention mechanism under different conditions defining bag positivity. As illustrated in Figure 1, the first experiment studies the case where both bag and instance class labels are known. We extend the application of ABMIL classifiers to the second setting which aims to explore how insights can be drawn from datasets where the threshold of positive instances, which define a positive bag, are unknown. Code will be available on GitHub https://github.com/NextPath-Lab/MIL-Instance-Scoring.

2.2.1 Proportion MNIST Bags Dataset

We refer to bagged MNIST datasets modified to accommodate the prev-MIL problem as Proportion MNIST Bags. As illustrated in Figure 1A, multiple datasets are generated by sampling MNIST digits across different values of $\tau \in \{0.0, 0.05, \ldots, 0.95, 1.0\}$. The number '9' is selected as the positive class instance for consistency with Ref. 10. Each dataset is approximately balanced in the number of positive and negative bags. When τ is not 0 or 1, bags are generated with an average size of 20 instances. Additionally, the distribution of positive instance proportions in positive and negative bags are centered at ± 0.15 from τ. Detailed composition of each dataset is available in Table 2 in Appendix A.

2.2.2 Prostate Cancer Large-scale DNA Organization Dataset

This dataset consists of 111 DNA-stoichiometric Feulgen-thionin stained prostate core needle biopsy images from 75 patients recruited from BC Cancer Agency affiliated hospitals from 1994-2012 with appropriate ethics permissions from the University of British Columbia and BC Cancer Agency. Clinicopathologic characteristics of the patients are summarized in Table 1. There are two defined groups of patients in this cohort: indolent PCa patients which refers to patients with Gleason score of 6 who did not progress under active surveillance, and aggressive PCa patients who was given a Gleason score of 9 and died within 2 years of consultation. Although distinguishing between these two outcomes is not clinically challenging, this study seeks to evaluate whether LDO features can capture prognostic information beyond conventional methods by first testing their predictive ability in this well-defined setting. To compute LDO features, nuclei in Feulgen-thionin stained biopsy sections were first segmented using a sequential attention U-Net model.[26,27] Delineated nuclei are then processed to calculate 140 LDO features describing nuclear morphology, stain intensity, and Haralick texture features. More details can be found in prior works,[20,25] and in Figure 5 in Appendix B. We preserve the semantics of LDO features by omitting the use of convolutional neural network (CNN) and multilayer perceptron (MLP) feature extractors in the ABMIL architecture to which we refer as ABMILite.

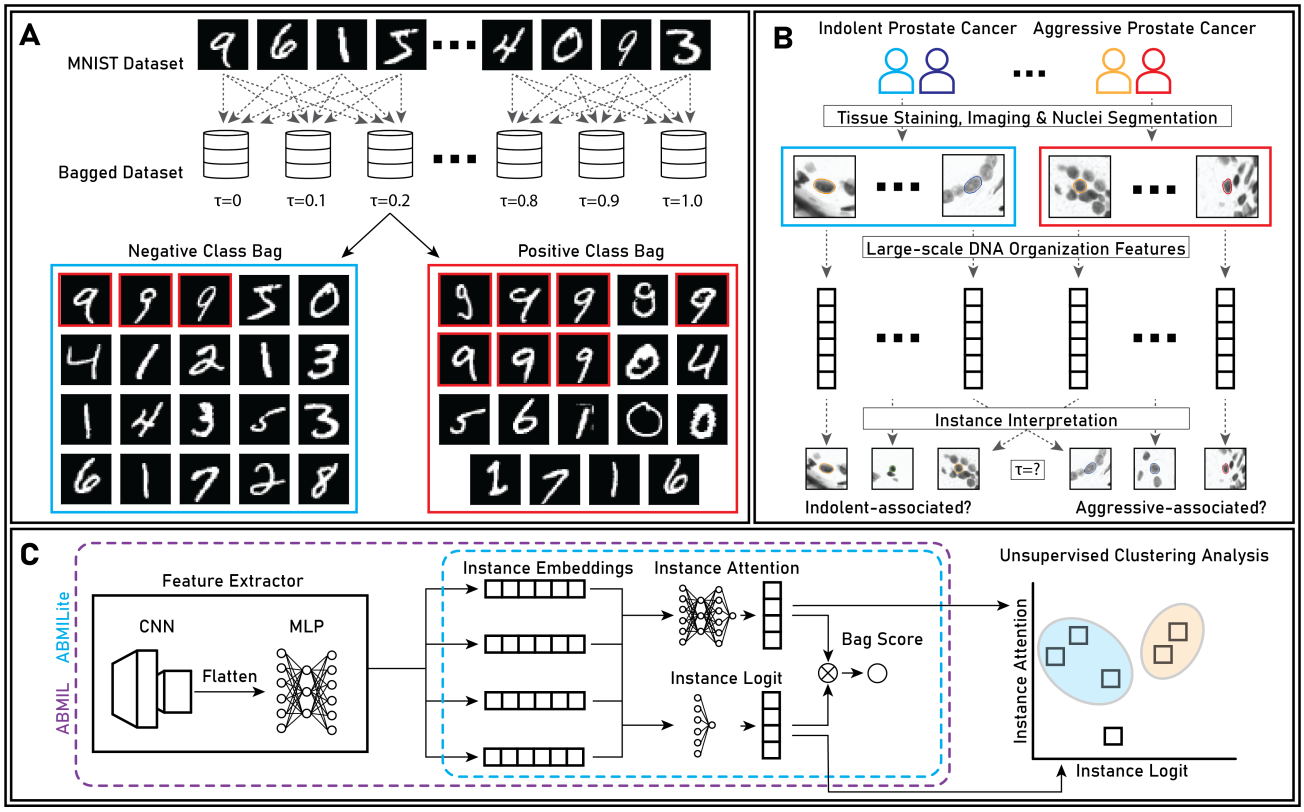

Figure 1. Illustration of the datasets used in this work. **(A)** shows how multiple datasets of bagged MNIST digits are created, based on different prevalence thresholds (τ) to classify a positive bag. The bags shown are examples from the dataset with $\tau = 0.2$ where up to 20% of negatively labeled bags may be the positive instance (MNIST digit 9). The **left** (blue) bag shows a negative bag with 15% composed by 9's, and the **right** (red) bag shows a positive bag with $\tilde{3}6.8\%$ 9's. **(B)** shows the prostate cancer data where large-scale DNA organization (LDO) features are computed from segmented images of Feulgen-thionin stained nuclei. As opposed to **(A)**, the τ value is unknown here, and instance classes must be inferred post-ABMIL training. Bottom row shows the attention-based multiple instance learning (ABMIL) architecture used for the MNIST-derived datasets, and ABMILite, a variant without feature extraction modules used for the LDO dataset.

Table 1. Clinicopathologic characteristics of prostate cancer patients from BC Cancer-affiliated hospitals

	Indolent Cancer	Aggressive Cancer
Patients	50	25
Biopsy images	38	73
Gleason score	6	9
Clinical Outcome	Active surveillance; no progression	Death (<2 years of consult.)
Mean nuclei count	6594	8769

2.3 Bag Classifier Distillation

We leverage unsupervised clustering approach in the distillation of instance contribution to bag-level classes. Post-model training, the output of the attention module and instance logit are analyzed together to create a 2 dimensional space visualizing the importance of the instance in bag-level prediction, and its resemblance to a positive class bag. We quantify the relationship between the measure of importance (attention) and positive class resemblance (instance logit) by studying their correlation, which is expected to be high if attention truly is a robust measure of instance class. Attention module outputs (attention logits) prior to SoftMax normalization are used in this analysis to remove the effect of bagging on attention weights.

Equation 2 shows that the bag-level logits can be expressed in terms of the instance scores by utilizing the

same weight matrices if a non-linearity is not applied. Thus, in a setting where $\tau = 0$, weights optimized for bag classification may also be an effective instance classifier. However, when $\tau > 0$, instance scores generated in this manner may not be well calibrated for instance classification since positive class instances are now introduced in negative bags. Thus, under the hypothesis that MIL classifiers can still rank instances under the prev-MIL condition, we leverage unsupervised clustering to identify subgroups of instances defined by their importance to bag prediction, and positivity score.

2.4 Model Description and Training

We train two models, ABMIL and ABMILite to analyze the Proportion MNIST Bags and PCa LDO dataset respectively. All training and analyses are performed using Python 3.11, scikit-learn (ver. 1.5.2) and PyTorch (ver. 2.4.1) using an NVIDIA RTX 3090 graphics processing unit (GPU).

2.4.1 ABMIL

The ABMIL architecture uses the implementation from Ref. 11 as shown in Figure 1C, where the classification head is applied to instance embeddings as opposed to bag-level embeddings. The ABMIL model is scaled down to 13,500 parameters and optimized for 25 epochs with a learning rate of 0.001 and the Adam optimizer. The models were trained with data bagged from the training set of the MNIST dataset from PyTorch (ver. 2.4.1) with no further splits for validation, and tested on data bagged from the MNIST test dataset. Model training is consistent on all MNIST-derived datasets.

2.4.2 ABMILite

The ABMILite architecture is the minimalist implementation of the general framework of attention-based MIL, without any additional feature extraction. As illustrated in Figure 1C, the ABMILite model consists of a two-layer MLP as the attention mechanism and a single linear classification layer with a total of 3,445 parameters. Omitting additional feature extraction modules preserves the semantics of the LDO features as inputs, enabling simpler interpretations of specific characteristics of nuclei which may be associated to aggressive PCa. We also apply L1 regularization during training to identify the most important features. ABMILite is trained with LDO features of nuclei from 56 patients, and tested on 19 patients, in a 3:1 train-test split.

3. RESULTS

3.1 Proportion MNIST Bags Classification

3.1.1 Negative Class Bags with Uniformly Sampled Digits

First, we demonstrate that the ABMIL classifier remains an effective classifier even when $\tau > 0$. This base case contains bags of MNIST digits where digits were sampled with a uniform probability for all digits. The expected proportion of the positive instance is 0.1. Positively labeled bags are further spiked with a mean proportion of '9' at 0.2, τ is set to 0.15. A total of 2053 training bags and 349 testing bags were generated from the training and testing sets of the MNIST dataset. The distribution of positive instance proportions is visualized in Figure 6 in Appendix C. The model achieved a balanced accuracy of 0.899 and 0.835 on the training and test set, respectively. As per Eq. 2, linear projections of instance features can be calculated using the weights of the ABMIL classification head. These projections are visualized in Fig. 2.

The number of clusters here was determined heuristically to be 3 to describe the high attention and instance logit cluster, low attention and low instance logit cluster and instances which fall in between. The cluster corresponding to the positive class instances was determined by selecting the cluster with the highest average instance logits. This method classified positive instances with a balanced accuracy of 0.9696 in comparison to using a standard sigmoid threshold of 0.5 on the instance logits which had a balanced accuracy of 0.9642.

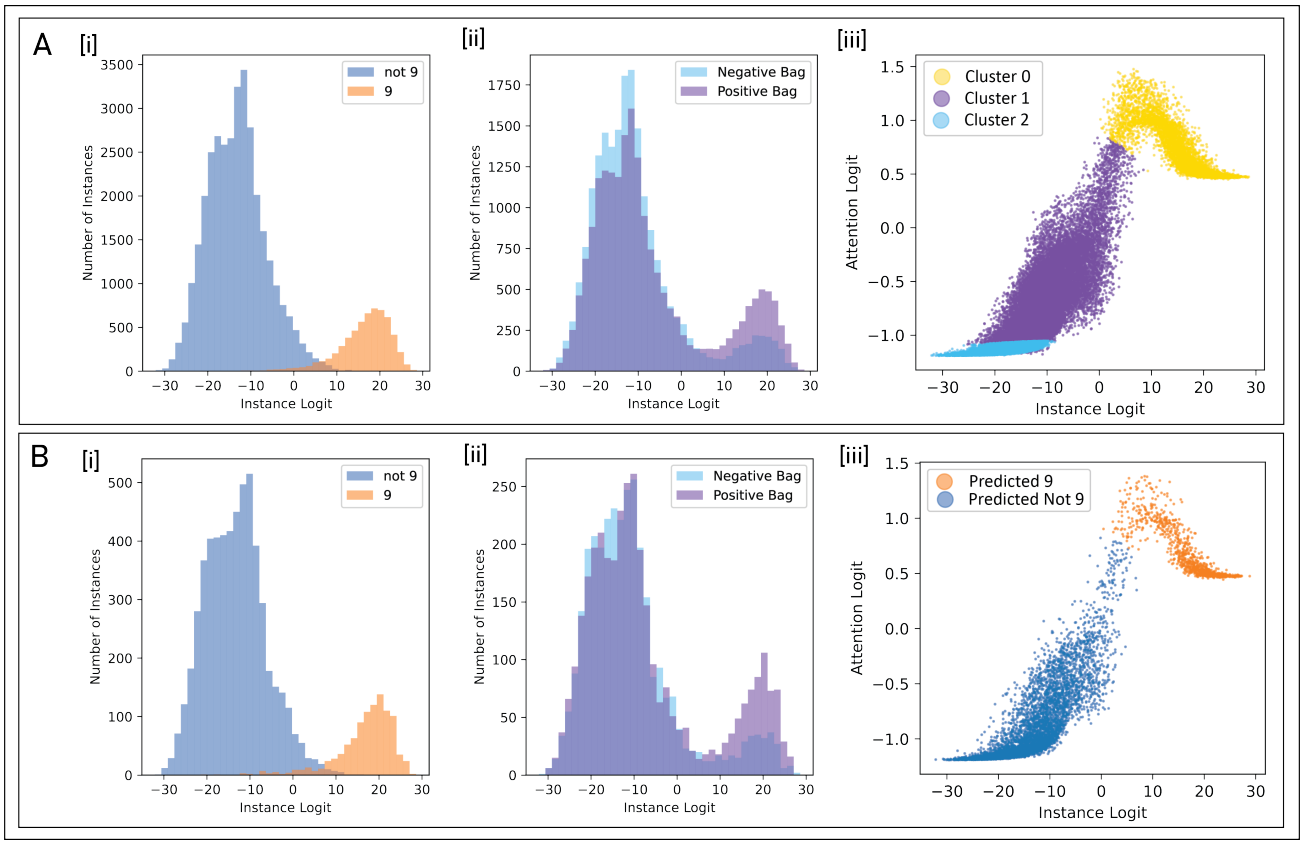

Figure 2. Results of the analysis of the case where negatively labeled bags are composed of randomly sampled MNIST digits. **(A)** shows results from training set, **(B)** shows results from the test set. Columns **[i]** and **[ii]** show the distribution of instance logits grouped by true instance label (not '9' vs is '9') and bag label respectively. **(A)[iii]** shows the results of unsupervised clustering. The Bayesian Variational Gaussian Mixture model trained on this set is applied to the test set in **(B)[iii]** to predict which instances are '9'.

3.1.2 Attention Behavior Across Varying Thresholds for Bag Positivity

Next we run the same experiment across 21 datasets where τ lies between 0 and 1, in 0.05 intervals. For each dataset, the ABMIL model is trained 5 times with different random seeds to account for variability across datasets. Unsupervised clustering for instance class interpretation is repeated for all permutations of ABMIL (trained with different seeds) for each dataset, as summarized in Figure 3A which displays the average balanced accuracy, F1 score and AUROC score respectively from [i-iii]. The performance of identifying positive class instances are compared with the classification head, using 0 as the threshold for logit values for binarization (corresponding to probability of 0.5 if sigmoid is applied). Aside from the dip in performance near τ of 0.6 to 0.75, this method performs as well as or slightly better than using the instance logit.

To study the change in behavior across different τ values, we illustrate four cases in Figure 3B[i-iv] which displays the 2D space of instance and attention logits for $\tau \in \{0.0, 0.25, 0.5, 0.75\}$ respectively. For smaller values of τ, the attention and instance logits show a high correlation, which disappears when τ reaches 0.5, and becomes a negative correlation at 0.75. The manner in how the correlation changes across τ values is shown in Figure 3C[i, ii], with Pearson and Spearman correlation respectively, to demonstrate both how linear and monotonic relationships. We also illustrate how τ affects bag classification performance in Figure 3C[iii] with the AUROC score.

3.2 BC Cancer PCa LDO Results

Following automatic feature filtration (described in **??**),[28] a total of 51 LDO features remained in the dataset, leaving the ABMILite model with a total of 3,445 parameters. The leave-one-out cross validation (LOOCV)

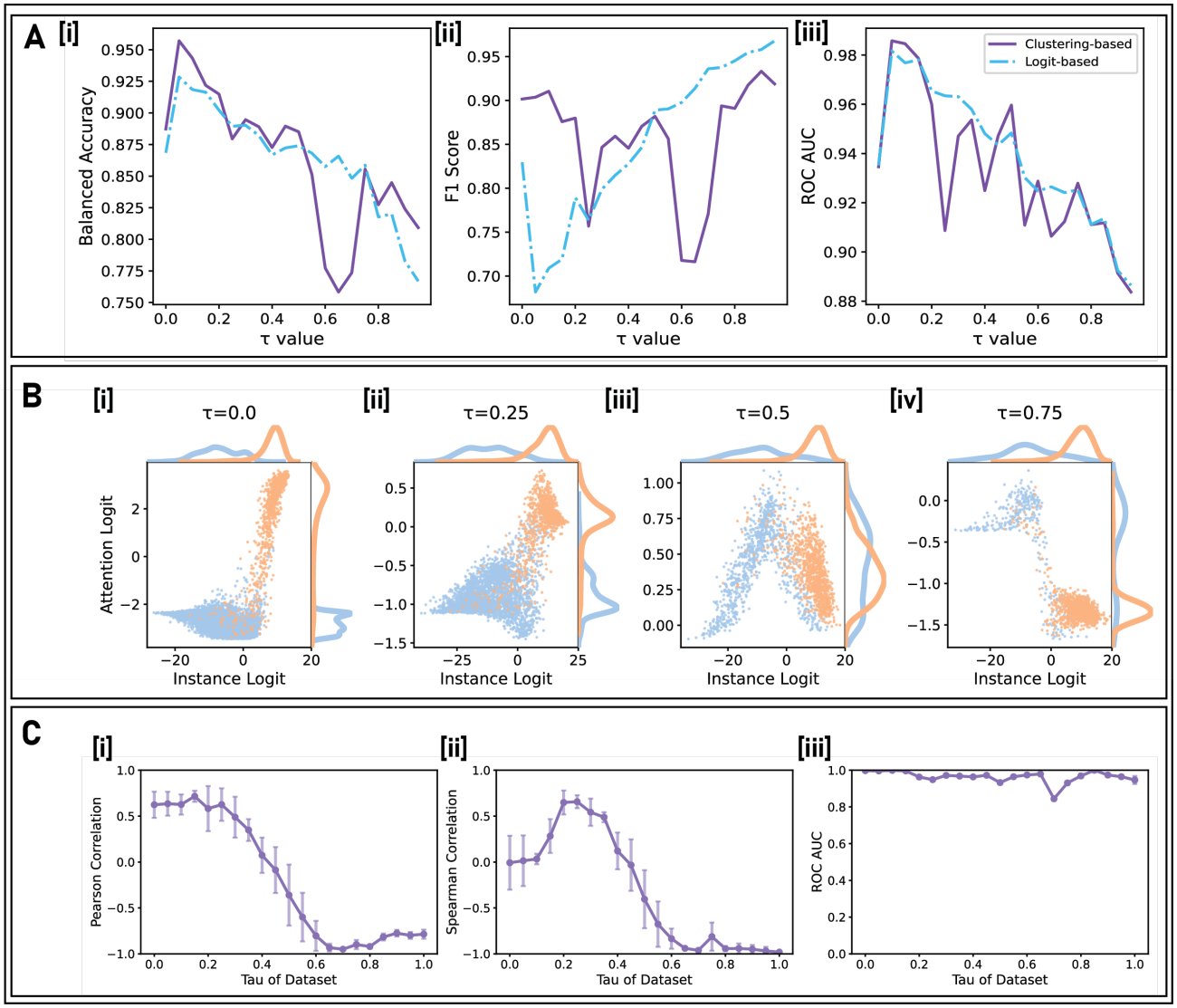

Figure 3. Correlation of τ and classification performance of ABMIL. The performance of positive instance classification (identifying '9's) for the unsupervised clustering method using Bayesian Gaussian Mixture model (in purple) and thresholding the instance logit at 0 (dotted blue) is shown in **(A)** for balanced accuracy, F1-score and ROC AUC respectively. **(B)** shows the distribution of attention and instance logit for four datasets illustrating how their correlation shifts from the traditional MIL setting, and for $\tau \in \{0.25, 0.5, 0.75\}$ respectively from [i-iv]. Colors represent the instance label, where blue shows non-'9' digits, and orange represents '9'. **(C)** illustrates the mean and standard deviation of the correlation of attention and instance logits across five different training seeds for all datasets. Both Pearson and Spearman correlation are shown to illustrate both monotonicity and linear relationships, and the bag classification AUROC is shown in [iii] to demonstrate the bag classification performance.

balanced accuracy for the training patients was 0.841, and 0.833 for the test patients. We report LOOCV metrics for the training set as we are limited by the small cohort size, and a validation set of one patient keeps the training set as consistent as possible across folds for a robust representation of its generalizability. For this analysis, the attention-weighted average of instance logits are referred to as the LDO score. The Bayesian Gaussian Mixture model identifies three clusters as intended, with one cluster corresponding to instances with both high attention and instance logit. This high attention, high instance logit cluster will be referred to as LDO+ nuclei, and the cluster with the lowest instance logits and a range of attention scores as LDO- nuclei. The LDO+ nuclei are likely those which are strongly associated to an aggressive PCa phenotype and further analysis

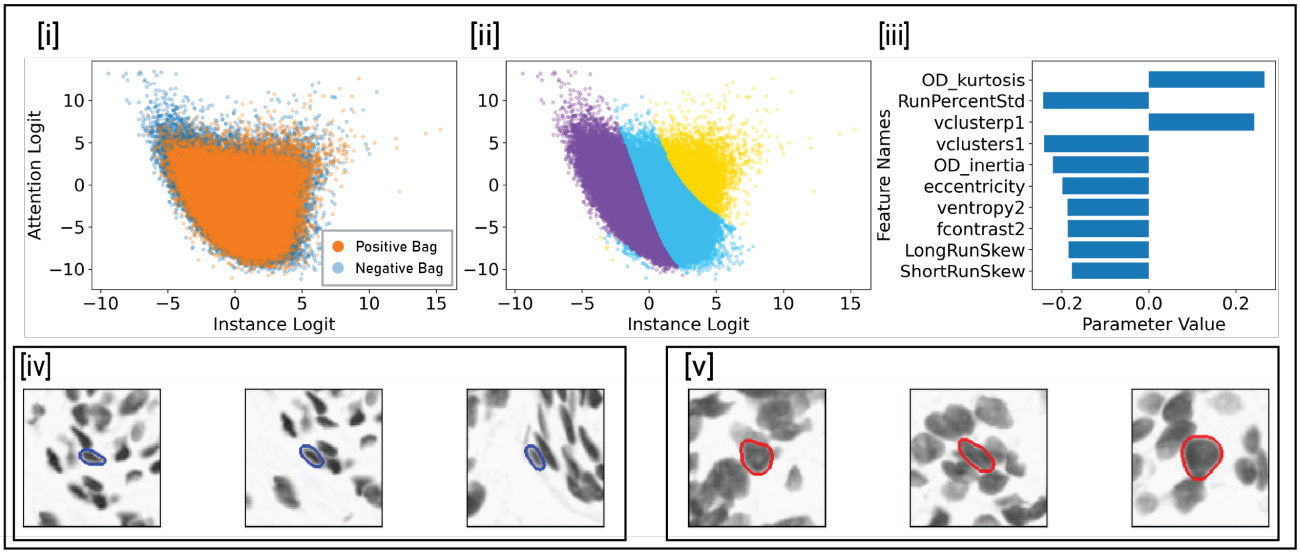

Figure 4. Results of the ABMILite prediction on the prostate cancer (PCa) LDO dataset. Negative bags refers to PCa patients with indolent cancer, positive bags refer to PCa patients with aggressive PCa. The distribution of attention and instance logits of the test patient group is displayed in **[i]**, grouped by patient outcome. **[ii]** shows the Bayesian Gaussian mixture clustering on the test set, which was trained on the training set. The yellow points represent LDO+ nuclei with high attention and instance logit. The 10 most important features are displayed in **[iii]**, where parameter value corresponds to its importance in the prediction due to L1 regularized training. Examples of LDO- with high attention logits and LDO+ nuclei are visualized in [iv,v] respectively.

of specific characteristics may shed light on the specific alterations which occur in these nuclei. The proportion of the cluster with the highest logit score was predictive of PCa aggressiveness on the test patients with a ROC AUC of 0.7949, where the average proportion of LDO+ nuclei in indolent PCa patients was 0.0309 and 0.0501 for aggressive PCa patients. Since the ABMILite model preserves the semantics of the LDO features and trained with L1 regularization, the parameter values correspond to its importance in patient prediction. The 10 most important features are displayed in Figure 4, all of which except *eccentricity* and *OD inertia* are measures of texture.

4. DISCUSSION

The objective of this study was to elucidate the manner in which the nature of MIL datasets may alter the attention mechanism from what is expected, and how that can be leveraged to extract further instance-level insights in pathology data. We demonstrated a simple method which can be used post-model training to generate subgroups from the attention and instance logit space, while most recent work aims to make ML models interpretable by design. This work takes a deeper look at the behavior of instance attention mechanisms, a crucial component of developing such interpretable models. The key finding of this work is that the nature of the dataset, and imprecision of measurements can affect how the attention mechanism behaves. By changing the prevalence of the positive class instance in bags of both positive and negative labels, we observe that the attention mechanism must learn to attend to non-positive instances. This is in contrast to the traditional MIL problem ($\tau = 0$), where the mechanism only needs to attend to positive instances, and maximizing its influence on the predictions. We thus conclude, that unless the dataset is curated in a manner that the negative bags are completely devoid of positive class instances, attention may not be a reliable indicator of instance class.

4.1 Bag Classifiers as Instance Classifiers

In the example of the MNIST-derived datasets, we observe that the ABMIL classification head is an effective predictor of instance class as shown in Figure 3A, although its performance degrades with greater τ. In Figure 3B, we see that the distribution of the positive instance logits remain stable across datasets with different positive

instance thresholds, and its the ability to distinguish negative instances which degrade. Particularly in the PCa LDO dataset, instance logits were not a reliable metric of classification where the its distribution was largely overlapped between patients with indolent and aggressive PCa. Thus, for deeper interpretability of instances, more dedicated model design or workflows may be necessary, such as INS which leverages prototype-based contrastive learning to generate instance pseudo-labels,[29] or Cluster-aware ABMIL which aims to learn different subgroups of instances to guide classification.[12] We are currently exploring the integration of such approaches in concurrent work. Such approaches are also advantageous since prev-MIL may change the expected behavior of the model, as we have seen in the attention mechanism of ABMIL. While this behavior can be demystified when different permutations of the dataset are available, with known values of τ, this is seldom the case. Placing constraints on the behavior of each model component thus reduces the degree of this alteration, making the interpretation of the behavior of each component more reliable. Such reliability is especially important in clinical settings, as clinical tools should aim to inform clinicians; not only through model predictions, but how the model arrived at such conclusion may also be informative. Such qualities are necessary to avoid over-reliance on ML-based tools and erosion of key clinical skills.[2,30,31]

4.2 Unsupervised Approach for Interpretability

Our approach conceptually is similar to that of prototype-based learning or cluster-aware learning in works such as Refs. 17 and 12 which aim to elucidate a feature 'signature' of subtypes of instances. In this work, we focus more on the out-of-the-box interpretation of ABMIL and apply unsupervised clustering post-hoc in contrast to others who have integrated it into the model. However, our results from applying Figure 3A demonstrate that post-hoc clustering may not always be the best performing approach as evidenced by the large drop in performance at τ between 0.6 and 0.75. This coincides with the drop in bag classification performance in C[iii] of the same figure, further reinforcing that instance-level interpretation of models optimized only on bag-level classification objectives may not be consistently reliable. An appropriate approach then, is to incorporate instance-level objective functions into the workflow such as in a prototype-based learning framework which is being investigated in concurrent work. That said, we did demonstrate post-hoc clustering may successfully identify instances useful for interpretation. The proportion of LDO+ nuclei, visualized in Figure 4[v], was predictive of aggressive PCa with an AUROC of 0.7949. This suggests LDO+ nuclei may have clinical relevance, although the definition of this category of nuclei may be noisier than if subgrouping nuclei were built into the ABMIL optimization objective. Further, 8 of the 10 most important features being measures of texture suggests there may be a difference in the molecular states that control chromatin behavior within epithelial cell nuclei between patients with indolent vs aggressive PCa. For instance, many biological phenomena are linked to a change in the chromatin organization which can be visualized on the WSI level. Among such phenomena are histone modifications such as hyperacetylation,[8] expression levels of transcription factors,[32,33] alterations to nuclear envelope and associated proteins among others.[9] The decreased expression for emerin, a nuclear lamina protein, for instance, is linked to cell mobility and cancer metastasis.[34] Alterations to nuclear structure which favors metastasis can be measured with LDO features, and we see that the eccentricity of nuclear shape happens to be in the 10 most important features to predict PCa. Alternatively, transcription factors such as Nuclear Factor I (NFIX) are tied closely to androgen receptor function,[32,33] which again is intimately linked to the progression of PCa. Specifically, Ref. 33 observed a close relationship between NFIX activity, chromatin accessibility, and the development of castration-resistant prostate cancer. Chromatin accessibility, or the distribution of hetero- and euchromatin can be quantified with Haralick texture features of nuclei stained with DNA-stoichiometric stains, as the intensity of each imaged pixel corresponds to the DNA amount in that spot.[20]

While our results suggest promising leads, it is not without limitations. While the imaging-based approach offers a more quantitative and reproducible approach to WSI analysis, the current LDO framework is limited to identifying molecular abnormalities that manifest as phenotypic changes in cell nuclei. This does not elucidate the specific cause for much molecular changes, where each different cause may have varying implications on patient survival as we have seen from the vast diversity of reasons why chromatin organization may be altered.[8,33,34] However, our framework may be an effective clinical test in which patients with particular patterns of nuclear changes can be flagged, and additional epigenetic, or genetic tests can be ordered at the physicians discretion.

5. CONCLUSION

We provided empirical evidence that attention weights are not indicators of instance class, especially in a prevalence-based MIL setting. Upon exploring how the behavior of instance attention may change, we observe negative class instances may also hold high attention values depending on the context of the dataset. The combined analysis of instance logits and attention in an unsupervised clustering approach demonstrated its utility in identifying positive class instances for the MNIST-derived datasets, and also to identify nuclei which are associated to aggressive PCa. The proportion of LDO+ nuclei identified through this approach was predictive of aggressive PCa with a AUROC of 0.7949. We outline the connection between characteristics of LDO+ nuclei and how they may provide clinicians with clinical insights beyond the model prediction. While this approach was effective in distilling instance-level insights which ABMIL had learned, there may be advantages to integrating instance subgrouping mechanisms into the model architecture itself which will be explored in future work. Future work will aim to (1) develop a simple ABMIL-derived architecture in which instance subgrouping is integrated, to enable more reliable instance-level interpretation, (2) associating LDO feature signatures to specific molecular alterations such as the increased expression of NFIX, and (3) validate LDO features as a biomarker of PCa aggressiveness on a larger cohort of PCa patients.

APPENDIX A. PROPORTION MNIST BAGS DATASET

Table 2. Dataset metadata across τ values. Train/Test balance indicates the proportion of positively labeled bags, demonstrating that all datasets are approximately balanced.

τ	Num. train bags	Num. test bags	Train balance	Test balance
0.00	3077	504	0.497	0.508
0.05	3077	498	0.494	0.520
0.10	2486	398	0.491	0.513
0.15	2050	342	0.495	0.526
0.20	1252	217	0.492	0.479
0.25	1225	216	0.496	0.463
0.30	1037	182	0.481	0.467
0.35	879	158	0.488	0.443
0.40	766	134	0.491	0.470
0.45	674	126	0.507	0.397
0.50	602	100	0.530	0.590
0.55	552	97	0.513	0.474
0.60	503	89	0.525	0.461
0.65	463	83	0.531	0.434
0.70	406	71	0.539	0.535
0.75	408	70	0.483	0.500
0.80	378	64	0.524	0.562
0.85	356	62	0.525	0.484
0.90	347	59	0.522	0.542
0.95	338	58	0.524	0.517
1.00	329	57	0.526	0.509

APPENDIX B. PROSTATE CANCER LARGE-SCALE DNA ORGANIZATION PIPELINE

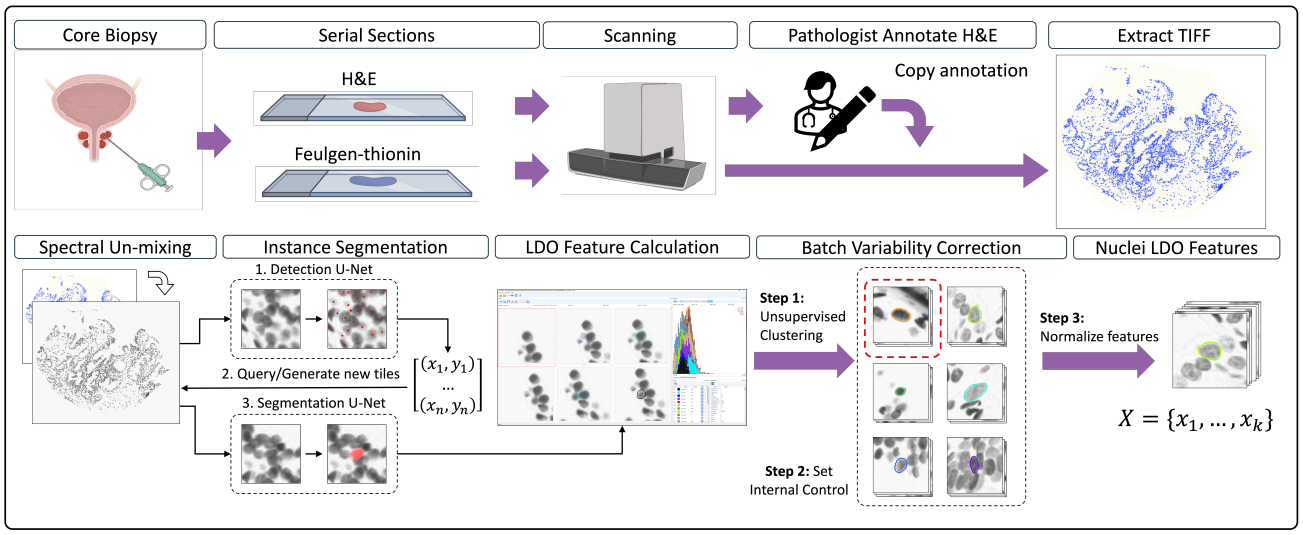

Figure 5. Overall imaging pipeline of Feulgen-thionin staining and nuclei analysis. Prostate core needle biopsies are taken and sectioned. Adjacent sections are stained with hematoxylin and eosin (H&E) and Feulgen-thionin. Slides are scanned at 20x on 3D Histech Ltd. Panoramic scanner. H&E stained WSI are annotated by a pathologist, marking regions of interest (ROI), subsequently copied onto the Feulgen scan. Annotated regions are extracted as TIFF images, and segmented with a sequential attention UNet. The first UNet marks geometric centers of nuclei which are then centered in 128x128 tiles, and segmented by the second UNet. 140 LDO features are calculated for each nuclear image. As Haralick features are affected by stain intensity, a heuristic cleaning step is applied to remove false positive and poorly segmented objects. Image intensities are normalized against a reference cell nuclei population within each TIFF image, identified by unsupervised clustering.

APPENDIX C. DISTRIBUTION OF MNIST DIGITS IN TRAIN/TEST SET

To ensure the Proportion MNIST Bags dataset is not too heavily influenced by a decrease in proportion of other MNIST digits, they are visualized here.

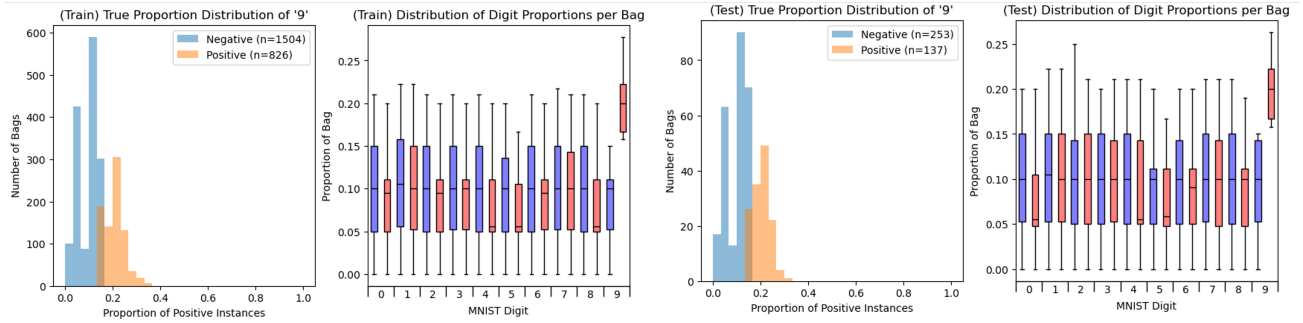

Figure 6. Illustration of the proportion MNIST bags dataset which follows our definition of the MIL problem in Eq. 1. with the number '9' as the positive instance. **Left:** a bag with a negative label, containing less than the τ threshold of 0.1 of positive instances. This bag depicts the expected frequency of each digit from 0-9 if digits are sampled equally, assuming the sampling probability for each digit is equal. **Right:** a bag with a positive label, containing the target number, 9, at a greater proportion than τ.

APPENDIX D. AUTOMATIC FEATURE FILTERING

Automatic feature filtration is done as in Ref. 28, with an additional Mutual Information score feature selection, wherein features with a mutual information score with the outcome label are removed. Mutual information is

calculated by taking nuclei-wise features and the patient's clinical outcome as nuclei pseudo-labels.

ACKNOWLEDGMENTS

We would like to acknowledge our funding and partner agencies. This work would not have been possible without the cooperation from the BC Cancer Foundation, funding support from the Canadian Institutes of Health Research (CIHR) under grant F15-00930 and Natural Science and Engineering Research Council of Canada (NSERC) under the Canadian Graduate Research Scholarship - Master's (CGRS-M) award from the University of British Columbia (#6563).

REFERENCES

[1] Goldenberg, S. L., Nir, G., and Salcudean, S. E., "A new era: artificial intelligence and machine learning in prostate cancer," *Nature Reviews Urology* **16**, 391–403 (July 2019).

[2] Azadi Moghadam, P., Bashashati, A., and Goldenberg, S. L., "Artificial Intelligence and Pathomics: Prostate Cancer," *The Urologic Clinics of North America* **51**, 15–26 (Feb. 2024).

[3] Canadian Cancer Statistics Advisory Committee, Canadian Cancer Society, Statistics Canada, and Public Health Agency of Canada, "Canadian Cancer Statistics 2025," statistical report, Toronto, ON (Nov. 2025).

[4] D'Amico, A. V., Whittington, R., Malkowicz, S. B., Schultz, D., Blank, K., Broderick, G. A., Tomaszewski, J. E., Renshaw, A. A., Kaplan, I., Beard, C. J., and Wein, A., "Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer," *JAMA* **280**, 969–974 (Sept. 1998).

[5] D'Amico, A. V., Moul, J., Carroll, P. R., Sun, L., Lubeck, D., and Chen, M.-H., "Cancer-specific mortality after surgery or radiation for patients with clinically localized prostate cancer managed during the prostate-specific antigen era," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **21**, 2163–2172 (June 2003).

[6] Klotz, L., Zhang, L., Lam, A., Nam, R., Mamedov, A., and Loblaw, A., "Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **28**, 126–131 (Jan. 2010).

[7] Yu, A., "Risk Stratification and Selection of Management Strategy for Localized Prostate Cancer," *Journal of the National Comprehensive Cancer Network* **22**, e245013 (May 2024).

[8] Watson, J. A., McKenna, D. J., Maxwell, P., Diamond, J., Arthur, K., McKelvey-Martin, V. J., and Hamilton, P. W., "Hyperacetylation in prostate cancer induces cell cycle aberrations, chromatin reorganization and altered gene expression profiles," *Journal of Cellular and Molecular Medicine* **14**, 1668–1682 (June 2010).

[9] Veltri, R. W. and Christudass, C. S., "Nuclear morphometry, epigenetic changes, and clinical relevance in prostate cancer," *Advances in Experimental Medicine and Biology* **773**, 77–99 (2014).

[10] Ilse, M., Tomczak, J. M., and Welling, M., "Attention-based Deep Multiple Instance Learning," (2018). Version Number: 4.

[11] Cai, L., Huang, S., Zhang, Y., Lu, J., and Zhang, Y., "Rethinking Attention-Based Multiple Instance Learning for Whole-Slide Pathological Image Classification: An Instance Attribute Viewpoint," (2024). Version Number: 1.

[12] Ko, S., Ando, Y., Kim, M., Park, N. J.-Y., Han, H., Park, J. Y., and Cho, J., "A cluster attention-based multiple instance learning network for enhancing histopathological image interpretation," *Computers in Biology and Medicine* **193**, 110353 (July 2025).

[13] Raciti, P., Sue, J., Retamero, J. A., Ceballos, R., Godrich, R., Kunz, J. D., Casson, A., Thiagarajan, D., Ebrahimzadeh, Z., Viret, J., Lee, D., Schüffler, P. J., DeMuth, G., Gulturk, E., Kanan, C., Rothrock, B., Reis-Filho, J., Klimstra, D. S., Reuter, V., and Fuchs, T. J., "Clinical Validation of Artificial Intelligence–Augmented Pathology Diagnosis Demonstrates Significant Gains in Diagnostic Accuracy in Prostate Cancer Detection," *Archives of Pathology & Laboratory Medicine* **147**, 1178–1185 (Oct. 2023).

[14] [*The Paige Prostate Suite: Assistive Artificial Intelligence for Prostate Cancer Diagnosis: Emerging Health Technologies*], CADTH Horizon Scans, Canadian Agency for Drugs and Technologies in Health, Ottawa (ON) (2024).

[15] Esteva, A., Feng, J., Van Der Wal, D., Huang, S.-C., Simko, J. P., DeVries, S., Chen, E., Schaeffer, E. M., Morgan, T. M., Sun, Y., Ghorbani, A., Naik, N., Nathawani, D., Socher, R., Michalski, J. M., Roach, M., Pisansky, T. M., Monson, J. M., Naz, F., Wallace, J., Ferguson, M. J., Bahary, J.-P., Zou, J., Lungren, M., Yeung, S., Ross, A. E., NRG Prostate Cancer AI Consortium, Kucharczyk, M., Souhami, L., Ballas, L., Peters, C. A., Liu, S., Balogh, A. G., Randolph-Jackson, P. D., Schwartz, D. L., Girvigian, M. R., Saito, N. G., Raben, A., Rabinovitch, R. A., Katato, K., Sandler, H. M., Tran, P. T., Spratt, D. E., Pugh, S., Feng, F. Y., and Mohamad, O., "Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials," *npj Digital Medicine* **5**, 71 (June 2022).

[16] Artera, Inc., "Arteraai prostate," (2025). FDA De Novo authorized test for prostate cancer prognostication.

[17] Yu, J.-G., Wu, Z., Ming, Y., Deng, S., Li, Y., Ou, C., He, C., Wang, B., Zhang, P., and Wang, Y., "Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images," *Medical Image Analysis* **85**, 102748 (Apr. 2023).

[18] Javed, S. A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., and Prakash, A., "Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology," (2022). Version Number: 2.

[19] Haralick, R. M., Shanmugam, K., and Dinstein, I., "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**, 610–621 (Nov. 1973).

[20] Whitaker, B. P., Stigliano, W. W., Carson, F. L., and Lynn, J. A., "Thionin Feulgen Stain for DNA (Ploidy) Quantitation by Image Analysis," *Journal of Histotechnology* **16**, 113–116 (June 1993).

[21] Zink, D., Fischer, A. H., and Nickerson, J. A., "Nuclear structure in cancer cells," *Nature Reviews. Cancer* **4**, 677–687 (Sept. 2004).

[22] Veltri, R. W., Miller, M. C., Partin, A. W., Coffey, D. S., and Epstein, J. I., "Ability to predict biochemical progression using Gleason score and a computer-generated quantitative nuclear grade derived from cancer cell nuclei," *Urology* **48**, 685–691 (Nov. 1996).

[23] Hveem, T. S., Kleppe, A., Vlatkovic, L., Ersvær, E., Wæhre, H., Nielsen, B., Kjær, M. A., Pradhan, M., Syvertsen, R. A., Nesheim, J. A., Liestøl, K., Albregtsen, F., and Danielsen, H. E., "Chromatin changes predict recurrence after radical prostatectomy," *British Journal of Cancer* **114**, 1243–1250 (May 2016).

[24] Zarei, N., Bakhtiari, A., Korbelik, J., Carraro, A., Keyes, M., Guillaud, M., and MacAulay, C., "Automated Region-based Prostate Cancer Cell Nuclei Localization. Part of a Prognostic Modality Tool for Prostate Cancer Patients," *Analytical and Quantitative Cytopathology and Histopathology* **38**, 59–69 (Apr. 2016).

[25] MacAulay, C., Keyes, M., Hayes, M., Lo, A., Wang, G., Guillaud, M., Gleave, M., Fazli, L., Korbelik, J., Collins, C., Keyes, S., and Palcic, B., "Quantification of large scale DNA organization for predicting prostate cancer recurrence," *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **91**, 1164–1174 (Dec. 2017).

[26] Macaulay, C. and Gallagher, P., "Sequential convolutional neural networks for nuclei segmentation," (July 2022).

[27] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D., "Attention U-Net: Learning Where to Look for the Pancreas," (May 2018). arXiv:1804.03999 [cs].

[28] Warkentin, M. T., Al-Sawaihey, H., Lam, S., Liu, G., Diergaarde, B., Yuan, J.-M., Wilson, D. O., Atkar-Khattra, S., Grant, B., Brhane, Y., Khodayari-Moez, E., Murison, K. R., Tammemagi, M. C., Campbell, K. R., and Hung, R. J., "Radiomics analysis to predict pulmonary nodule malignancy using machine learning approaches," *Thorax* **79**, 307–315 (Mar. 2024).

[29] Qu, L., Ma, Y., Luo, X., Wang, M., and Song, Z., "Rethinking Multiple Instance Learning for Whole Slide Image Classification: A Good Instance Classifier is All You Need," (2023). Version Number: 2.

[30] Harada, Y., Katsukura, S., Kawamura, R., and Shimizu, T., "Effects of a Differential Diagnosis List of Artificial Intelligence on Differential Diagnoses by Physicians: An Exploratory Analysis of Data from a Randomized Controlled Study," *International Journal of Environmental Research and Public Health* **18**, 5562 (May 2021).

[31] Cabitza, F., Campagner, A., and Sconfienza, L. M., "Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading," *Health Information Science and Systems* **9**, 8 (Dec. 2021).

[32] Grabowska, M. M., Elliott, A. D., DeGraff, D. J., Anderson, P. D., Anumanthan, G., Yamashita, H., Sun, Q., Friedman, D. B., Hachey, D. L., Yu, X., Sheehan, J. H., Ahn, J.-M., Raj, G. V., Piston, D. W., Gronostajski, R. M., and Matusik, R. J., "NFI transcription factors interact with FOXA1 to regulate prostate-specific gene expression," *Molecular Endocrinology (Baltimore, Md.)* **28**, 949–964 (June 2014).

[33] Poluben, L., Nouri, M., Liang, J., Chen, S., Varkaris, A., Ersoy-Fazlioglu, B., Voznesensky, O., Lee, I. I., Qiu, X., Cato, L., Seo, J.-H., Freedman, M. L., Sowalsky, A. G., Lack, N. A., Corey, E., Nelson, P. S., Brown, M., Long, H. W., Russo, J. W., and Balk, S. P., "Increased nuclear factor I-mediated chromatin access drives transition to androgen receptor splice variant dependence in prostate cancer," *Cell Reports* **44**, 115089 (Jan. 2025).

[34] Liddane, A. G. and Holaska, J. M., "The Role of Emerin in Cancer Progression and Metastasis," *International Journal of Molecular Sciences* **22**, 11289 (Oct. 2021).